

Recent Advances in Convolutional Neural Networks

Jiuxiang Gu*, Zhenhua Wang*, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, and Gang Wang, *Member, IEEE*

Abstract—In the last few years, deep learning has led to very good performance on a variety of problems, such as visual recognition, speech recognition and natural language processing. Among different types of deep neural networks, convolutional neural networks have been most extensively studied. Due to the lack of training data and computing power in early days, it is hard to train a large high-capacity convolutional neural network without overfitting. After the rapid growth in the amount of the annotated data and the recent improvements in the strengths of graphics processor units (GPUs), the research on convolutional neural networks has been emerged swiftly and achieved state-of-the-art results on various tasks. In this paper, we provide a broad survey of the recent advances in convolutional neural networks. Besides, we also introduce some applications of convolutional neural networks in computer vision.

Index Terms—Convolutional Neural Network, Deep learning.

I. INTRODUCTION

CONVOLUTIONAL Neural Network (CNN) is a well-known deep learning architecture inspired by the natural visual perception mechanism of the living creatures. In 1959, Hubel & Wiesel [1] found that cells in animal visual cortex are responsible for detecting light in receptive fields. Inspired by this discovery, Kunihiro Fukushima proposed the neocognitron in 1980 [2], which could be regarded as the predecessor of CNN. In 1990, LeCun *et al.* [3] published the seminal paper establishing the modern framework of CNN, and later improved it in [4]. They developed a multi-layer artificial neural network called LeNet-5 which could classify handwritten digits. Like other neural networks, LeNet-5 has multiple layers and can be trained with the backpropagation algorithm [5]. It can obtain effective representations of the original image, which makes it possible to recognize visual patterns directly from raw pixels with little-to-none preprocessing. However, due to the lack of large training data and computing power at that time, LeNet-5 can not perform well on more complex problems, *e.g.*, large-scale image and video classification.

Since 2006, many methods have been developed to overcome the difficulties encountered in training deep CNNs. Most notably, Krizhevsky *et al.* propose a classic CNN architecture and show significant improvements upon previous methods on the image classification task. The overall architecture of their method, *i.e.*, AlexNet [6], is similar to LeNet-5 but with a deeper structure. With the success of AlexNet, several works are proposed to improve its performance. Among

them, four representative works are ZFNet [7], VGGNet [8], GoogleNet [9] and ResNet [10]. From the evolution of the architectures, a typical trend is that the networks are getting deeper, *e.g.*, ResNet, which won the champion of ILSVRC 2015, is about 20 times deeper than AlexNet and 8 times deeper than VGGNet. By increasing depth, the network can better approximate the target function with increased non-linearity and get better feature representations. However, it also increases the complexity of the network, which makes the network be more difficult to optimize and easier to get overfitting. Along this way, various methods are proposed to deal with these problems in various aspects. In this paper, we try to give a comprehensive review of recent advances and give some thorough discussions.

In the following sections, we identify broad categories of works related to CNN. We first give an overview of the basic components of CNN in Section II. Then, we introduce some recent improvements on different aspects of CNN including convolutional layer, pooling layer, activation function, loss function, regularization and optimization in Section III and introduce the fast computing techniques in Section IV. Next, we discuss some typical applications of CNN including image classification, object detection, object tracking, pose estimation, text detection and recognition, visual saliency detection, action recognition and scene labeling in Section V. Finally, we conclude this paper in Section VI.

II. BASIC CNN COMPONENTS

There are numerous variants of CNN architectures in the literature. However, their basic components are very similar. Take the famous LeNet-5 [4] as an example (see Fig. 1), it consists of three types of layers, namely convolutional, pooling, and fully-connected layers.

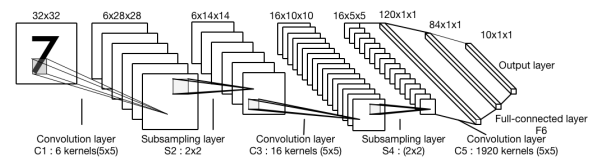


Fig. 1: The architecture of LeNet-5 network, which works well on digit classification task.

The convolutional layer aims to learn feature representations of the inputs. As shown in Fig. 1, convolution layer is composed of several convolution kernels which are used to compute different feature maps. Specifically, each neuron of a feature map is connected to a neighborhood of neurons in the previous layer. Such a neighborhood is referred to as the

JX. Gu, ZH. Wang, J. Kuen, LY. Ma, A. Shahroudy, B. Shuai, T. Liu, XX. Wang, G. Wang are with the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore (e-mail: jgu004@ntu.edu.sg; zhwang.me@gmail.com; jasonkuen@ntu.edu.sg; lyima@ntu.edu.sg; amir3@ntu.edu.sg; bshuai001@ntu.edu.sg; LIUT0016@e.ntu.edu.sg; wangxx@ntu.edu.sg; WangGang@ntu.edu.sg).

* equal contribution.

neuron's receptive field in the previous layer. The new feature map can be obtained by first convolving the input with a learned kernel and then applying an element-wise nonlinear activation function on the convolved results. By using several different kernels, the complete new feature maps are obtained. Mathematically, the feature value at location (i, j) in the k -th feature map of l -th layer, $z_{i,j,k}^l$, is calculated by:

$$z_{i,j,k}^l = \mathbf{w}_k^l \mathbf{x}_{i,j}^l + b_k^l \quad (1)$$

where \mathbf{w}_k^l and b_k^l are the weight vector and bias term of the k -th filter of the l -th layer respectively, and $\mathbf{x}_{i,j}^l$ is the input patch centered at location (i, j) of the l -th layer. Note that the kernel \mathbf{w}_k^l that generates the feature map $\mathbf{z}_{i,j,k}^l$ is shared. Such a weight sharing mechanism has several advantages such as it can reduce the model complexity and makes the network easier to train. The activation function introduces nonlinearities to CNN, which are desirable for multi-layer networks to detect nonlinear features. Let $a(\cdot)$ denote the nonlinear activation function. The activation value $a_{i,j,k}^l$ of convolutional feature $z_{i,j,k}^l$ can be computed as:

$$a_{i,j,k}^l = a(z_{i,j,k}^l) \quad (2)$$

Typical activation functions are sigmoid, tanh and ReLU [11]. The pooling layer aims to achieve shift-invariance by reducing the resolution of the feature maps. It is usually placed between two convolutional layers. Each feature map of a pooling layer is connected to its corresponding feature map of the preceding convolutional layer. Denoting the pooling function as $\text{pool}(\cdot)$, for each feature map $a_{i,j,k}^l$ we have:

$$y_{i,j,k}^l = \text{pool}(a_{m,n,k}^l), \forall (m, n) \in \mathcal{R}_{ij} \quad (3)$$

where \mathcal{R}_{ij} is a local neighbourhood around location (i, j) . The typical pooling operations are average pooling [12] and max pooling [13–15]. By stacking several convolutional and pooling layers, we could extract more abstract feature representations.

After several convolutional and pooling layers, there may be one or more fully-connected layers which aim to perform high level reasoning [7], [8], [16]. They take all neurons in the previous layer and connect them to every single neuron of current layer to generate global semantic information. The outputs of the last fully-connected layer will be fed to an output layer. For classification tasks, softmax operator is commonly used [6]. Another commonly used method is SVM, which can be combined with CNN features to solve different classification tasks [17]. Let θ denote all the parameters of a CNN (e.g., the weight vectors and bias terms). The optimum parameters for a specific task can be obtained by minimizing an appropriate loss function defined on that task. Suppose we have N desired input-output relations $\{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}); n \in [1, \dots, N]\}$, where $\mathbf{x}^{(n)}$ is the n -th input data, $\mathbf{y}^{(n)}$ is its corresponding target label and $\mathbf{o}^{(n)}$ is the output of CNN. The loss of CNN can be calculated as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \ell(\theta; \mathbf{y}^{(n)}, \mathbf{o}^{(n)}) \quad (4)$$

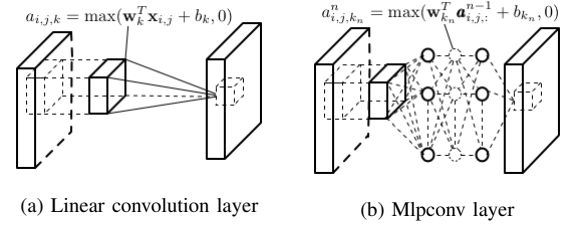


Fig. 2: The comparison of linear convolution layer and mlpconv layer.

Training a CNN is a problem of global optimization. By minimizing the loss function, we can find the best fitting set of parameters. Stochastic gradient descent is a common solution for optimizing CNN network [18–20].

III. IMPROVEMENTS ON CNNs

There have been various improvements on CNNs since the success of AlexNet in 2012. In this section, we describe the major improvements on CNNs from six aspects: convolutional layer, pooling layer, activation function, loss function, regularization and optimization.

A. Convolutional layer

Convolution filter in basic CNNs is a generalized linear model (GLM) for the underlying local image patch. It works well for abstraction when instances of latent concepts are linearly separable. Here we introduce two works which aim to enhance its representation ability.

1) *Network in network*: Network In Network (NIN) is a general network structure proposed by Lin *et al.* [21]. It replaces the linear filter of the convolutional layer by a micro network, e.g., multilayer perceptron convolution (mlpconv) layer in the paper, which makes it capable of approximating more abstract representations of the latent concepts. The overall structure of NIN is the stacking of such micro networks. Fig. 2 shows the difference between linear convolutional layer and mlpconv layer. Formally, the feature map of convolution layer (with nonlinear activation function, e.g., ReLU [11]) is computed as:

$$a_{i,j,k} = \max(\mathbf{w}_k^T \mathbf{x}_{i,j} + b_k, 0) \quad (5)$$

where $a_{i,j,k}$ is the activation value of k -th feature map at location (i, j) , $\mathbf{x}_{i,j}$ is the input patch centered at location (i, j) , \mathbf{w}_k and b_k are weight vector and bias term of the k -th filter. As a comparison, the computation performed by mlpconv layer is formulated as:

$$a_{i,j,k_n}^n = \max(\mathbf{w}_{k_n}^T \mathbf{a}_{i,j,:}^{n-1} + b_{k_n}, 0) \quad (6)$$

where $n \in [1, N]$, N is the number of layers in the mlpconv layer, $\mathbf{a}_{i,j,:}^0$ is equal to $\mathbf{x}_{i,j}$. Therefore the mlpconv layer can also be regarded as cascaded cross channel parametric pooling on a normal convolutional layer.

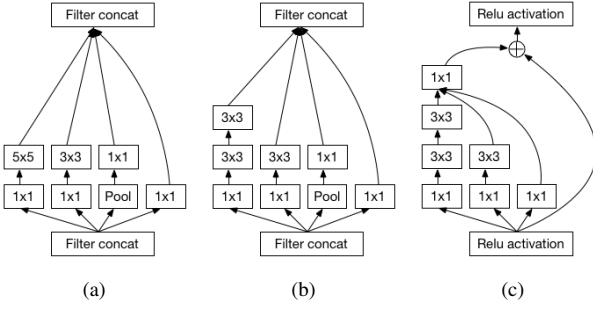


Fig. 3: (a) The inception module used in [9]. (b) The improved inception module used in [22] where each 5×5 convolution is replaced by two 3×3 convolutions. (c) The Inception-ResNet-A module used in [23].

2) *Inception module*: Inception module is introduced by Szegedy *et al.* [9] which can be seen as a logical culmination of NIN. [9] uses variable filter sizes to capture different visual patterns of different sizes, and approximates the optimal sparse structure by the inception module. Specifically, inception module consists of one pooling operation and three types of convolution operations (see Fig. 3(a)). 1×1 convolutions are placed before 3×3 and 5×5 convolutions as dimension reduction modules, which allow for increasing the depth and width of CNN without increasing the computational complexity. With the help of inception module, the network parameters can be dramatically reduced to 5 millions which are much less than those of AlexNet (60 millions) and ZFNet (75 millions). In their later paper [22], to find high performance networks with a relatively modest computation cost, they suggest the representation size should gently decrease from inputs to outputs as well as spatial aggregation can be done over lower dimensional embeddings without much loss in representational power. The optimal performance of the network can be reached by balancing the number of filters per layer and the depth of the network. Inspired by the ResNet [10], their latest work [23] combines the inception architecture with shortcut connections (see Fig. 3(c)). They find that shortcut connections can significantly accelerate the training of inception networks. Their Inception-v4 model architecture (with 75 trainable layers) that ensembles three residual and one Inception-v4 can achieve 3.08% top-5 error rate on the validation dataset of ILSVRC 2012.

B. Pooling layer

Pooling is an important concept of CNN. It lowers the computational burden by reducing the number of connections between convolutional layers. In this section, we introduce some recent pooling methods used in CNNs.

1) *L_p Pooling*: L_p pooling is a biologically inspired pooling process modelled on complex cells [24], [25]. It has been theoretically analysed in [26], [27], which suggest that L_p pooling provides better generalization than max pooling. L_p pooling can be represented as :

$$y_{i,j,k} = \left[\sum_{(m,n) \in \mathcal{R}_{ij}} (a_{m,n,k})^p \right]^{1/p} \quad (7)$$

where $y_{i,j,k}$ is the output of the pooling operator at location (i, j) in k -th feature map, and $a_{m,n,k}$ is the feature value at location (m, n) within the pooling region \mathcal{R}_{ij} in k -th feature map. Specially, when $p = 1$, L_p corresponds to average pooling, and when $p = \infty$, L_p reduces to max pooling.

2) *Mixed Pooling*: Inspired by random Dropout [16] and DropConnect [28], Yu *et al.* [29] propose a mixed pooling method which is the combination of max pooling and average pooling. The function of mixed pooling can be formulated as follows:

$$y_{i,j,k} = \lambda \max_{(m,n) \in \mathcal{R}_{ij}} a_{m,n,k} + (1 - \lambda) \frac{1}{|\mathcal{R}_{ij}|} \sum_{(m,n) \in \mathcal{R}_{ij}} a_{m,n,k} \quad (8)$$

where λ is a random value being either 0 or 1 which indicates the choice of either using average pooling or max pooling. During forward propagation process, λ is recorded and will be used for the backpropagation operation. Experiments in [29] show that mixed pooling can better address the overfitting problems and it performs better than max pooling and average pooling.

3) *Stochastic pooling*: Stochastic pooling [30] is a dropout-inspired pooling method. Instead of picking the maximum value within each pooling region as max pooling does, stochastic pooling randomly picks the activations according to a multinomial distribution, which ensures that the non-maximal activations of feature maps are also possible to be utilized. Specifically, stochastic pooling first computes the probabilities p for each region \mathcal{R}_j by normalizing the activations within the region, i.e., $p_i = a_i / \sum_{k \in \mathcal{R}_j} (a_k)$. After obtaining the distribution $P(p_1, \dots, p_{|\mathcal{R}_j|})$, we can sample from the multinomial distribution based on p to pick a location l within the region, then set the pooled activation as $y_j = a_l$, where $l \sim P(p_1, \dots, p_{|\mathcal{R}_j|})$. Compared with max pooling, stochastic pooling can avoid overfitting due to the stochastic component.

4) *Spectral pooling*: Spectral pooling [31] performs dimensionality reduction by cropping the representation of input in frequency domain. Given an input feature map $\mathbf{x} \in \mathbb{R}^{m \times m}$, suppose the dimension of desired output feature map is $h \times w$, spectral pooling first computes the discrete Fourier transform (DFT) of the input feature map, then crops the frequency representation by maintaining only the central $h \times w$ submatrix of the frequencies, and finally uses inverse DFT to map the approximation back into spatial domain. Compared with max pooling, the linear low-pass filtering operation of spectral pooling can preserve more information for the same output dimensionality. Meanwhile, it also does not suffer from the sharp reduction in output map dimensionality exhibited by other pooling methods. What is more, the process of spectral pooling is achieved by matrix truncation, which makes it capable of being implemented with little computational cost in CNNs (e.g., [32]) that employ FFT for convolution kernels.

5) *Spatial pyramid pooling*: Spatial pyramid pooling (SPP) is introduced by He *et al.* [33]. The key advantage of SPP is that it can generate a fixed-length representation regardless of the input sizes. SPP pools input feature map in local spatial bins with sizes proportional to the image size, resulting in a fixed number of bins. This is different from the sliding window

pooling in the previous deep networks, where the number of sliding windows depends on the input size. By replacing the last pooling layer with SPP, they propose a new SPP-net which is able to deal with images with different sizes.

6) *Multi-scale Orderless Pooling*: Inspired by [34], Gong *et al.* [35] use multi-scale orderless pooling (MOP) to improve the invariance of CNNs without degrading their discriminative power. They extract deep activation features for both the whole image and local patches of several scales. The activations of the whole image is the same as those of previous CNNs, which aim to capture the global spatial layout information. The activations of local patches are aggregated by VLAD encoding [36], which aim to capture more local, fine-grained details of the image as well as enhancing invariance. The new image representation is obtained by concatenating the global activations and the VLAD features of the local patch activations.

C. Activation function

A proper activation function significantly improves the performance of a CNN for a certain task. In this section, we introduce the recent used activation functions in CNNs.

1) *ReLU*: Rectified linear unit (ReLU) [11] is one of the most notable non-saturated activation functions. The ReLU activation function is defined as:

$$a_{i,j,k} = \max(z_{i,j,k}, 0) \quad (9)$$

where $z_{i,j,k}$ is the input of the activation function at location (i, j) on the k -th channel. ReLU is a piecewise linear function which prunes the negative part to zero and retains the positive part (see Fig. 4(a)). The simple $\max(\cdot)$ operation of ReLU allows it to compute much faster than sigmoid or tanh activation functions, and it also induces the sparsity in the hidden units and allows the network to easily obtain sparse representations. It has been shown that deep networks can be trained efficiently using ReLU even without pre-training [6]. Even though the discontinuity of ReLU at 0 may hurt the performance of backpropagation, many works have shown that ReLU works better than sigmoid and tanh activation functions empirically [37–39].

2) *Leaky ReLU*: A potential disadvantage of ReLU unit is that it has zero gradient whenever the unit is not active. This may cause units that do not active initially never active as the gradient-based optimization will not adjust their weights. Also, it may slow down the training process due to the constant zero gradients. To alleviate this problem, Mass *et al.* introduce leaky ReLU (LReLU) [37] which is defined as:

$$a_{i,j,k} = \max(z_{i,j,k}, 0) + \lambda \min(z_{i,j,k}, 0) \quad (10)$$

where λ is a predefined parameter in range $(0, 1)$. Compared with ReLU, Leaky ReLU compresses the negative part rather than mapping it to constant zero, which makes it allow for a small, non-zero gradient when the unit is not active.

3) *Parametric ReLU*: Rather than using a predefined parameter in Leaky ReLU, *e.g.*, λ in Eq.(10), He *et al.* [40] propose Parametric Rectified Linear Unit (PReLU) which adaptively learns the parameters of the rectifiers in order to

improve accuracy. Mathematically, PReLU function is defined as:

$$a_{i,j,k} = \max(z_{i,j,k}, 0) + \lambda_k \min(z_{i,j,k}, 0) \quad (11)$$

where λ_k is the learned parameter for the k -th channel. As PReLU only introduces a very small number of extra parameters, *e.g.*, the number of extra parameters is the same as the number of channels of the whole network, there is no extra risk of overfitting and the extra computational cost is negligible. It also can be simultaneously trained with other parameters by backpropagation.

4) *Randomized ReLU*: Another variant of Leaky ReLU is Randomized Leaky Rectified Linear Unit (RReLU) [41]. In RReLU, the parameters of negative parts are randomly sampled from a uniform distribution in training, and then fixed in testing (see Fig. 4(c)). Formally, RReLU function is defined as:

$$a_{i,j,k}^{(n)} = \max(z_{i,j,k}^{(n)}, 0) + \lambda_k^{(n)} \min(z_{i,j,k}^{(n)}, 0) \quad (12)$$

where $z_{i,j,k}^{(n)}$ denotes the input of activation function at location (i, j) on the k -th channel of n -th example, $\lambda_k^{(n)}$ denotes its corresponding sampled parameter, and $a_{i,j,k}^{(n)}$ denotes its corresponding output. It could reduce overfitting due to its randomized nature. [41] also evaluates ReLU, LReLU, PReLU and RReLU on standard image classification task, and concludes that incorporating a non-zero slop for negative part in rectified activation units could consistently improve the performance.

5) *ELU*: [42] introduces Exponential Linear Unit (ELU) which enables faster learning of deep neural networks and leads to higher classification accuracies. Like ReLU, LReLU, PReLU and RReLU, ELU avoids the vanishing gradient problem by setting the positive part to identity. In contrast to ReLU, ELU has a negative part which is beneficial for fast learning.

Compared with LReLU, PReLU and RReLU which also have unsaturated negative parts, ELU employs a saturation function as negative part. As the saturation function will decrease the variation of the units if deactivated, it makes ELU more robust to noise. The function of ELU is defined as:

$$a_{i,j,k} = \max(z_{i,j,k}, 0) + \min(\lambda(e^{z_{i,j,k}} - 1), 0) \quad (13)$$

where λ is a predefined parameter for controlling the value to which an ELU saturate for negative inputs.

6) *Maxout*: Maxout [43] is an alternative nonlinear function that takes the maximum response across multiple channels at each spatial position. As stated in [43], the maxout function is defined as: $a_{i,j,k} = \max_{k \in [1, K]} z_{i,j,k}$, where $z_{i,j,k}$ is the k -th channel of the feature map. It is worth noting that maxout enjoys all the benefits of ReLU since ReLU is actually a special case of maxout, *e.g.*, $\max(\mathbf{w}_1^T \mathbf{x} + b_1, \mathbf{w}_2^T \mathbf{x} + b_2)$ where \mathbf{w}_1 is a zero vector and b_1 is zero. Besides, maxout is particularly well suited for training with Dropout.

7) *Probout*: [44] proposes a probabilistic variant of maxout called probout. They replace the maximum operation in maxout with a probabilistic sampling procedure. Specifically, they first define a probability for each of the k linear units as: $p_i = e^{\lambda z_i} / \sum_{j=1}^k e^{\lambda z_j}$, where λ is a hyperparameter for controlling the variance of the distribution. Then, they pick

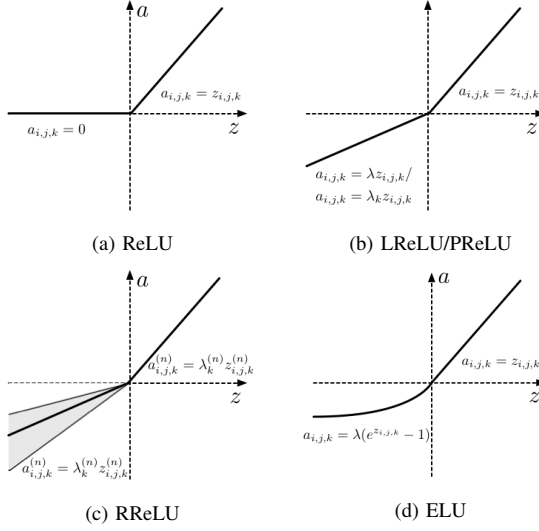


Fig. 4: The comparison among ReLU, LReLU, PReLU, RReLU and ELU. For Leaky ReLU, λ is empirically pre-defined. For PReLU, λ_k is learned from training data. For RReLU, $\lambda_k^{(n)}$ is a random variable which is sampled from a given uniform distribution in training and keeps fixed in testing. For ELU, λ is empirically predefined.

one of the k units according to a Multinomial distribution $\{p_1, \dots, p_k\}$ and set the activation value to be the value of picked unit. In order to incorporate with dropout, they actually re-define the probabilities as:

$$\hat{p}_0 = 0.5, \quad \hat{p}_i = e^{\lambda z_i} / (2 \cdot \sum_{j=1}^k e^{\lambda z_j}) \quad (14)$$

The activation function is then sampled as:

$$a_i = \begin{cases} 0 & \text{if } i = 0 \\ z_i & \text{else} \end{cases} \quad (15)$$

where $i \sim \text{Multinomial}\{\hat{p}_0, \dots, \hat{p}_k\}$. Probout can achieve the balance between preserving the desirable properties of maxout units and improving their invariance properties. However, in testing process, probout is computationally expensive than maxout due to the additional probability calculations.

D. Loss function

It is important to choose an appropriate loss functions for a specific task. We introduce three representative ones in this subsection: softmax loss, hinge loss, and contrastive loss.

1) *Softmax loss*: Softmax loss is a commonly used loss function which is essentially a combination of multinomial logistic loss and softmax. Given a training set $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}); i \in 1, \dots, N, \mathbf{y}^{(i)} \in 1, \dots, K\}$, where $\mathbf{x}^{(i)}$ is the i -th input image patch, and $\mathbf{y}^{(i)}$ is its target class label among the K classes. The prediction $a_j^{(i)}$ of j -th class for i -th input is transformed with the softmax function: $p_j^{(i)} = e^{a_j^{(i)}} / \sum_{l=1}^K e^{a_l^{(i)}}$. Softmax turns the predictions into non-negative values and normalizes them to get a probability distribution over classes. Such

probabilistic predictions are used to compute the multinomial logistic loss, *i.e.*, the softmax loss, as follows:

$$\mathcal{L} = -\frac{1}{N} \left[\sum_{i=1}^N \sum_{j=1}^K 1\{\mathbf{y}^{(i)} = j\} \log p_j^{(i)} \right] \quad (16)$$

2) *Hinge loss*: Hinge loss is usually used to train large margin classifiers such as Support Vector Machine (SVM). The hinge loss function of a multi-class SVM is defined in Eq.(17), where \mathbf{w} is the weight vector of classifier and $\mathbf{y}^{(i)} \in [1, \dots, K]$ indicates its correct class label among the K classes.

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K [\max(0, 1 - \delta(\mathbf{y}^{(i)}, j) \mathbf{w}^T \mathbf{x}_i)]^p \quad (17)$$

where $\delta(\mathbf{y}^{(i)}, j) = 1$ if $\mathbf{y}^{(i)} = j$, otherwise $\delta(\mathbf{y}^{(i)}, j) = -1$. Note that if $p = 1$, Eq.(17) is Hinge-Loss (L_1 -Loss), while if $p = 2$, it is the Squared Hinge-Loss (L_2 -Loss) [45]. The L_2 -Loss is differentiable and imposes a larger loss for point which violates the margin comparing with L_1 -Loss. [17] investigates and compares the performance of softmax with L_2 -SVMs in deep networks. The results on MNIST [46] demonstrate the superiority of L_2 -SVM over softmax.

3) *Contrastive loss*: Contrastive loss is commonly used to train Siamese network [47] which is a weakly-supervised scheme for learning a similarity measure from pairs of data instances labelled as matching or non-matching. The contrastive loss is usually defined for every layer $l \in [1, \dots, L]$ and the backpropagations for the loss of individual layers are performed at the same time. Given a pair of data $(\mathbf{x}_\alpha, \mathbf{x}_\beta)$, let (z_α^l, z_β^l) denote the output pair of layer l , the contrastive loss \mathcal{L} is defined as:

$$\mathcal{L} = \frac{1}{2L} \sum_{l=1}^L (y) d^l + (1 - y) \max(m - d^l, 0) \quad (18)$$

where $d^l = \|\mathbf{z}_\alpha^l - \mathbf{z}_\beta^l\|_2^2$ is the Euclidean distance between \mathbf{z}_α^l and \mathbf{z}_β^l , and m is a margin parameter affecting non-matching pairs. If $(\mathbf{x}_\alpha, \mathbf{x}_\beta)$ is a matching pair, then $y = 1$. Otherwise, $y = 0$. This loss function is also referred to as a single margin parameter loss function. Lin *et al.* [48] find that such a single margin loss function causes a dramatic drop in retrieval results when fine-tuning the network on all pairs. Meanwhile, the performance is better retained when fine-tuning only on non-matching pairs. This indicates that the handling of matching pairs in the loss function is responsible for the drop. While the recall rate on non-matching pairs alone is stable, handling the matching pairs is the main reason for the drop of recall rate. To solve this problem, they propose a double margin loss function which adds another margin parameter to affect the matching pairs:

$$\mathcal{L} = \frac{1}{2L} \sum_{l=1}^L (y) \max(d^l - m_1, 0) + (1 - y) \max(m_2 - d^l, 0) \quad (19)$$

E. Regularization

Overfitting is an unneglectable problem in deep CNNs, which can be effectively reduced by regularization. In the

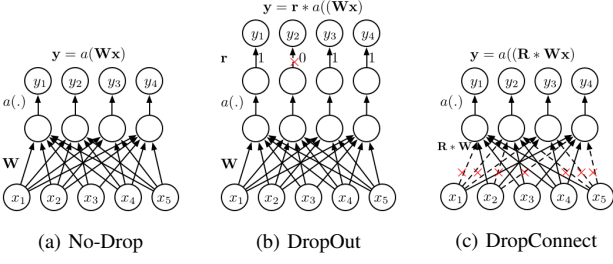


Fig. 5: The illustration of No-Drop network, Dropout network and DropConnect network.

following subsection, we introduce two effective regularization techniques: Dropout [16], [49], [50] and DropConnect [28].

1) *Dropout*: Dropout is first introduced by Hinton *et al.* [16], and it has been proven to be very effective in reducing overfitting. In [16], they apply Dropout to fully-connected layers. The output of Dropout is $\mathbf{y} = \mathbf{r} * \mathbf{a}(\mathbf{W}^T \mathbf{x})$, where $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$ is the input to fully-connected layer, \mathbf{W} (of size $d \times n$) is a weight matrix, and \mathbf{r} is a binary vector of size d whose elements are independently drawn from a Bernoulli distribution with parameter p , i.e. $r_i \sim \text{Bernoulli}(p)$. Dropout can prevent the network from becoming too dependent on any one (or any small combination) of neurons, and can force the network to be accurate even in the absence of certain information. Several methods have been proposed to improve Dropout. [49] proposes a fast Dropout method which can perform fast Dropout training by sampling from or integrating a Gaussian approximation. [50] proposes an adaptive Dropout method, where the Dropout probability for each hidden variable is computed using a binary belief network that shares parameters with the deep network. In [51], they find that applying standard Dropout before 1×1 convolutional layer generally increases training time but does not prevent overfitting. Therefore, they propose a new Dropout method called SpatialDropout, which extends the Dropout value across the entire feature map. This new Dropout method works well especially when the training data size is small.

2) *DropConnect*: DropConnect [28] takes the idea of Dropout a step further. Instead of randomly setting the outputs of neurons to zero, DropConnect randomly sets the elements of weight matrix \mathbf{W} to zero. The output of DropConnect is given by $\mathbf{y} = \mathbf{a}((\mathbf{R} * \mathbf{W})\mathbf{x})$, where $R_{ij} \sim \text{Bernoulli}(p)$. Additionally, the biases are also masked out during the training process. Fig. 5 illustrates the differences among No-Drop, Dropout and DropConnect networks.

F. Optimization

In this subsection, we discuss some key techniques for optimizing CNNs.

1) *Weights initialization*: Deep CNN has a huge amount of parameters and its loss function is non-convex [52], which makes it very difficult to train. To achieve a fast convergence in training and avoid the vanishing gradient problem, a proper network initialization is one of the most important prerequisites [53], [54]. The bias parameters can be initialized to zero,

while the weight parameters should be initialized carefully to break the symmetry among hidden units of the same layer. If the network is not properly initialized, e.g., each layer scales its input by k , the final output will scale the original input by k^L where L is the number of layers. In this case, value of $k > 1$ leads to extremely large values of output layers while value of $k < 1$ leads a diminishing output value and gradients. Krizhevsky *et al.* [6] initialize the weights of their network from a zero-mean Gaussian distribution with standard deviation 0.01 and set the bias terms of the second, fourth and fifth convolutional layers as well as all the fully-connected layers to constant one. Another famous random initialization method is “Xavier”, which is proposed in [55]. They pick the weights from a Gaussian distribution with zero mean and a variance of $2/(n_{\text{in}} + n_{\text{out}})$, where n_{in} is the number of neurons feeding into it, and n_{out} is the number of neurons the result is fed to. Thus “Xavier” can automatically determine the scale of initialization based on the number of input and output neurons, and keep the signal in a reasonable range of values through many layers. One of its variants in Caffe¹ uses the n_{in} -only variant, which makes it much easier to implement. “Xavier” initialization method is later extended by [40] to account for the rectifying nonlinearities, where they derive a robust initialization method that particularly considers the ReLU nonlinearity. Their method, allows for the training of extremely deep models (e.g., [9]) to converge while the “Xavier” method [55] cannot.

Independently, [56] shows that orthonormal matrix initialization works much better for linear networks than Gaussian initialization, and it also works for networks with nonlinearities. [53] extends [56] to an iterative procedure. Specifically, it proposes a layer-sequential unit-variance process scheme which can be viewed as an orthonormal initialization combined with batch normalization (see Section III-F3) performed only on the first mini-batch. It is similar to batch normalization as both of them take a unit variance normalization procedure. Differently, it uses ortho-normalization to initialize the weights which helps to efficiently de-correlate layer activities. Such an initialization technique has been applied to [57–59] with a remarkable increase in performance.

2) *Stochastic gradient descent*: The backpropagation algorithm [60] is the standard training method which uses gradient descent to update the parameters. Many gradient descent optimization algorithms have been proposed [61–64]. Standard gradient descent algorithm updates the parameters θ of the objective $\mathcal{L}(\theta)$ as $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} E[\mathcal{L}(\theta_t)]$, where $E[\mathcal{L}(\theta_t)]$ is the expectation of $\mathcal{L}(\theta)$ over the full training set and η is the learning rate. Instead of computing $E[\mathcal{L}(\theta_t)]$, stochastic gradient descent (SGD) [18], [19] estimates the gradients on the basis of a single randomly picked example $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ from the training set:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_{\theta} \mathcal{L}(\theta_t; \mathbf{x}^{(t)}, \mathbf{y}^{(t)}) \quad (20)$$

In practice, each parameter update in SGD is computed with respect to a mini-batch as opposed to a single example. This could help to reduce the variance in the parameter update and

¹<https://github.com/BVLC/caffe>

can lead to more stable convergency. The convergence speed is controlled by the learning rate η_t . A common method is to use a constant learning rate that gives stable convergence in the initial stage, and then reduce the learning rate as the convergence slows down.

Parallelized SGD methods [20], [65], [66] improve SGD to be suitable for parallel, large-scale machine learning. Unlike standard (synchronous) SGD in which the training will be delayed if one of the machines is slow, these parallelized methods use the asynchronous mechanism so that no other optimizations will be delayed except for the one on the slowest machine. Jeffrey Dean *et al.* [67] use another asynchronous SGD procedure called Downpour SGD to speed up the large-scale distributed training process on clusters with many CPUs. There are also some works that use asynchronous SGD with multiple GPUs. [68] basically combines asynchronous SGD with GPUs to accelerate the training time by several times compared to training on a single machine. [69] also uses multiple GPUs to asynchronously calculate gradients and update the global model parameters, which achieves 3.2 times of speedup on 4 GPUs compared to training on a single GPU.

3) *Batch Normalization*: Data normalization is usually the first step of data preprocessing. Global data normalization transforms all the data to have zero-mean and unit variance. However, as the data flows through a deep network, the distribution of input to internal layers will be changed, which will lose the learning capacity and accuracy of the network. [70] proposes an efficient method called Batch Normalization (BN) to partially alleviate this phenomenon. It accomplishes the so called covariate shift problem by a normalization step that fixes the means and variances of layer inputs where estimation of mean and variance are computed after each mini-batch rather than entire training set. Suppose the layer to normalize has a d dimensional input, *i.e.*, $\mathbf{x} = [x_1, x_2, \dots, x_d]^T$. We first normalize the k -th dimension as follows:

$$\hat{x}_k = (x_k - \mu_B) / \sqrt{\delta_B^2 + \epsilon} \quad (21)$$

where μ_B and δ_B^2 are the mean and variance of mini-batch respectively, and ϵ is a constant value. To enhance the representation ability, the normalized input \hat{x}_k is further transformed into:

$$y_k = \text{BN}_{\gamma, \beta}(x_k) = \gamma \hat{x}_k + \beta \quad (22)$$

where γ and β are learned parameters. Batch normalization has many advantages compared with global data normalization. Firstly, it reduces internal covariant shift. Secondly, BN reduces the dependence of gradients on the scale of the parameters or of their initial values, which gives a beneficial effect on the gradient flow through the network. This enables the use of higher learning rate without the risk of divergence. Furthermore, BN regularizes the model, and thus reduces the need for Dropout. Finally, BN makes it possible to use saturating nonlinear activation functions without getting stuck in the saturated model.

4) *Shortcut connections*: As mentioned above, the vanishing gradient problem of deep CNNs can be alleviated by normalized initialization [6] and BN [70]. Although these methods

successfully prevent deep neural networks from overfitting, they also introduce difficulties in optimizing the networks, resulting worse performances than shallower networks [40], [55], [56]. Such an optimization problem suffered by deeper CNNs is regarded as the degradation problem.

Inspired by Long Short Term Memory (LSTM) recurrent neural networks [71] which use gate functions to determine how much of a neuron's activation value to transform or just pass through. [72] proposes highway networks which enable the optimization of networks with virtually arbitrary depth. The output of their network is given by $\mathbf{y} = \phi(\mathbf{x}, \mathbf{W}_H) \cdot \tau(\mathbf{x}, \mathbf{W}_T) + \mathbf{x} \cdot (1 - \tau(\mathbf{x}, \mathbf{W}_T))$, where $\tau(\cdot)$ is the transform gate and $\phi(\cdot)$ is usually an affine transformation followed by a non-linear activation function (in general it may take other forms). This gating mechanism forces the layer's inputs and outputs to be of the same size and allows highway networks with tens or hundreds of layers to be trained efficiently. The outputs of gates vary significantly with the input examples, demonstrating that the network does not just learn a fixed structure, but dynamically routes data based on specific examples.

Independently, Residual Nets (ResNets) [10] share the same core idea that worked in LSTM units. Instead of employing learnable weights for neuron-specific gating, the shortcut connections in ResNets are not gated and untransformed input is directly propagated to the output which brings less parameters. With residual block, activation of any deeper unit can be written as the sum of the activation of a shallower unit and a residual function. This also implies that gradients can be directly propagated to shallower units, which makes deep ResNets much easier to be optimized than the original mapping function and more efficient to train very deep nets. This is in contrast to usual feedforward networks, where gradients are essentially a series of matrix-vector products, that may vanish as networks grow deeper. In their later paper [73], they use a set of experiments to show that identity shortcut connections are easiest for the network to learn. They also find that bringing BN forward can normalize the input signals to all weight layers which can avoid messing up information flow and performs considerably better than using BN after addition. In their comparisons, the residual net with BN + ReLU pre-activation gets higher accuracies than their previous ResNets [10]. The latest Inception paper [23] also reports that training is accelerated and performance is improved by using identity skip connections across Inception modules. In all of the ResNets [10], Highway [72] and Inception networks [23], we can notice that the same inputs do travel through paths. There seems to be a pretty clear trend of using shortcut connections to help train very deep networks. By contrast, [74] decreases the depth and increases the width of ResNets, and achieves impressive results on CIFAR10, CIFAR100 and SVHN. However, their claims are not validated on the large scale image classification task on Imagenet dataset².

²<http://www.image-net.org>

IV. FAST PROCESSING OF CNNs

With the increasing challenges in the computer vision and machine learning tasks, the models of deep neural networks get more and more complex. These powerful models require more data for training in order to avoid overfitting. Meanwhile, the big training data also brings new challenges such as how to train the networks in a feasible amount of times. In this section, we introduce some fast processing methods of CNNs.

A. FFT

Mathieu *et al.* [32] carry out the convolutional operation in the Fourier domain with FFTs. Using FFT-based methods has many advantages. Firstly, the Fourier transformations of filters can be reused as the filters are convolved with multiple images in a mini-batch. Secondly, the Fourier transformations of the output gradients can be reused when backpropagating gradients to both filters and input images. Finally, summation over input channels can be performed in the Fourier domain, so that inverse Fourier transformations are only required once per output channel per image. There have already been some GPU-based libraries developed to speed up the training and testing process, such as cuDNN [75] and fbfft [76]. However, using FFT to perform convolution needs additional memory to store the feature maps in the Fourier domain, since the filters must be padded to be the same size as the inputs. This is especially costly when the striding parameter is larger than 1, which is common in many state-of-art networks, such as the early layers in [77] and [9].

B. Matrix Factorization

Low-rank matrix factorization has been exploited in a variety of contexts to improve the optimization problems. Given an $m \times n$ matrix A of rank r , there exists a factorization $A = B \times C$ where B is an $m \times r$ full column rank matrix and C is an $r \times n$ full row rank matrix. Thus, we can replace A by B and C . To reduce the parameters of A by a fraction p , it is essential to ensure that $mr + rn < pmn$, i.e., the rank of A should satisfy that $r < pmn/(m+n)$. To this end, [78] applies the low-rank matrix factorization to the final weight layer in a deep CNN, resulting about 30-50% speedup in training time at little loss in accuracy. Similarly, [79] applies singular value decomposition on each layer of a deep CNN to reduce the model size by 71% with less than 1% relative accuracy loss. Inspired by [80] which demonstrates the redundancy in the parameters of deep neural networks, Denton *et al.* [81] and Jaderberg *et al.* [82] independently investigate the redundancy within the convolutional filters and develop approximations to reduced the required computations.

C. Vector quantization

Vector Quantization (VQ) is a method for compressing densely connected layers to make CNN models smaller. Similar to scalar quantization where a large set of numbers is mapped to a smaller set [83], VQ quantizes groups of numbers together rather than addressing them one at a time. In 2013, Denil *et al.* [80] demonstrate the presence of redundancy in

neural network parameters, and use VQ to significantly reduce the number of dynamic parameters in deep models. Gong *et al.* [84] investigate the information theoretical vector quantization methods for compressing the parameters of CNNs, and they obtain parameter prediction results similar to those of [80]. They also find that VQ methods have a clear gain over existing matrix factorization methods, and among the VQ methods, structured quantization methods such as product quantization work significantly better than other methods (*e.g.*, residual quantization [85], scalar quantization [86]).

V. APPLICATIONS OF CNNs

In this section, we introduce some recent works that apply CNNs to achieve state-of-the-art performance, including image classification, object tracking, pose estimation, text detection, visual saliency detection, action recognition and scene labeling.

A. Image Classification

CNNs have been applied in image classification for a long time [87–90]. Compared with other methods, CNNs can achieve better classification accuracy on large scale datasets [6], [8], [91], [92] due to their capability of joint feature and classifier learning. The breakthrough of large scale image classification comes in 2012. Krizhevsky *et al.* [6] develop the AlexNet and achieve the best performance in ILSVRC 2012. After the success of AlexNet, several works have made significant improvements in classification accuracy by either reducing filter size [7] or expanding the network depth [8], [9].

Building a hierarchy of classifiers is a common strategy for image classification with a large number of classes [93]. The work of [94] is one of the earliest attempts to introduce category hierarchy in CNN, in which a discriminative transfer learning with tree-based priors is proposed. They use a hierarchy of classes for sharing information among related classes in order to improve performance for classes with very few training examples. Similarly, [95] builds a tree structure to learn fine-grained features for subcategory recognition. [96] proposes a training method that grows a network not only incrementally but also hierarchically. In their method, classes are grouped according to similarities and are self-organized into different levels. [97] introduces a hierarchical deep CNNs (HD-CNNs) by embedding deep CNNs into a category hierarchy. They decompose the classification task into two steps. The coarse category CNN classifier is first used to separate easy classes from each other, and then those more challenging classes are routed downstream to fine category classifiers for further prediction. This architecture follows the coarse-to-fine classification paradigm and can achieve lower error at the cost of affordable increase of complexity.

Subcategory classification is another rapidly growing sub-field of image classification. There are already some fine-grained image datasets (such as Flower [98], Birds [99], [100], Dogs [101], Cars [102] and Shoes [103]). Using object part information is beneficial for fine-grained classification.

Generally, the accuracy can be improved by localizing important parts of objects and representing their appearances discriminatively. Along this way, Branson *et al.* [104] propose a method which detects parts and extracts CNN features from multiple pose-normalized regions. Part annotation information is used to learn a compact pose normalization space. They also build a model that integrates lower-level feature layers with pose-normalized extraction routines and higher-level feature layers with unaligned image features to improve the classification accuracy. Zhang *et al.* [105] propose a part-based R-CNN which can learn whole-object and part detectors. They use selective search [106] to generate the part proposals, and apply non-parametric geometric constraints to more accurately localize parts. Lin *et al.* [107] incorporate part localization, alignment, and classification into one recognition system which is called Deep LAC. Their system is composed of three sub-networks: localization sub-network is used to estimate the part location, alignment sub-network receives the location as input and performs template alignment [108], and classification sub-network takes pose aligned part images as input to predict the category label. They also propose a value linkage function to link the sub-networks and make them work as a whole in training and testing.

As can be noted, all the above mentioned methods make use of part annotation information for supervised training. However, these annotations are not easy to collect and these systems have difficulty in scaling up and to handle many types of fine grained classes. To avoid this problem, some researchers propose to find localized parts or regions in an unsupervised manner. Krause *et al.* [109] use the ensemble of localized learned feature representations for fine-grained classification, they use co-segmentation and alignment to generate parts, and then compare the appearance of each part and aggregate the similarities together. In their latest paper [110]. They combine co-segmentation and alignment in a discriminative mixture to generate parts for facilitating fine-grained classification. [111] applies visual attention in CNN for fine-grained classification. Their classification pipeline is composed of three types of attentions: the bottom-up attention proposes candidate patches, the object-level top-down attention selects relevant patches of a certain object, and the part-level top-down attention localizes discriminative parts. These attentions are combined to train domain specific networks which can help to find foreground object or object parts and extract discriminative features. [112] proposes bilinear models for fine-grained image classification. The recognition architecture consists of two feature extractors. The outputs of two feature extractors are multiplied using outer product at each location of the image, and are pooled to obtain an image descriptor.

B. Object detection

Object detection has been a long-standing and important problem in computer vision [113–115]. Generally, the difficulties mainly lie in how to accurately and efficiently localize objects in images or video frames. The use of CNNs for detection and localization [116–119] can be traced back to 1990s. However, due to the lack of training data and limited

processing resources, the progress of CNN-based object detection is slow before 2012. Since 2012, the huge success of CNNs in ImageNet challenge [6] rekindles interest in CNN-based object detection [120–126]. In some early works [116], [118], [127], they use the sliding window based approaches to densely evaluate the CNN classifier on windows sampled at each location and scale. Since there are usually hundreds of thousands of candidate windows in a image, these methods suffer from highly computational cost, which makes them unsuitable to be applied on large-scale dataset, *e.g.*, Pascal VOC [92], ImageNet [91] and MSCOCO [128].

Recently, object proposal based methods attract a lot of interests and are widely studied in the literature [106], [129–132]. These methods usually exploit fast and generic measurements to test whether a sampled window is a potential object or not, and further pass the output object proposals to more sophisticated detectors to determine whether they are background or belong to a specific object class. One of the most famous object proposal based CNN detector is Region-based CNN (R-CNN) [133]. R-CNN uses Selective Search (SS) [106] to extract around 2000 bottom-up region proposals that are likely to contain objects. Then, these region proposals are warped to a fix size (227×227), and a pre-trained CNN is used to extract features on them. Finally, a binary SVM classifier is used for detection.

R-CNN yields a significant performance boost. However, its computational cost is still high since the time consuming CNN feature extractor will be performed for each region separately. To deal with this problem, some recent works propose to share the computation in feature extraction [8], [77], [133], [134]. OverFeat [77] computes CNN features from an image pyramid for localization and detection. Hence the computation can be easily shared between overlapping windows. Spatial pyramid pooling network (SPP net) [135] is a pyramid-based version of R-CNN [133], which introduces a SPP layer to relax the constraint that input images must have a fixed size. Unlike R-CNN [133], SPP net extracts the feature maps from the entire image only once, and then applies spatial pyramid pooling on each candidate window to get a fixed-length representation. A drawback of SPP net is that its training procedure is a multi-stage pipeline, which makes it impossible to train the CNN feature extractor and SVM classifier jointly to further improve the accuracy. Fast RCNN [136] improves SPP net by using an end-to-end training method. All network layers can be updated during fine-tuning, which simplifies the learning process and improves detection accuracy. Later, Faster R-CNN [136] introduces a region proposal network (RPN) for object proposals generation and achieves further speed-up. Beside R-CNN based methods, [137] proposes a multi-region and semantic segmentation-aware model for object detection. They integrate the combined features on an iterative localization mechanism as well as a box-voting scheme after non-max suppression. [138] treats the object detection problem as an iterative classification problem. It predicts an accurate object boundary box by aggregating quantized weak directions from their detection network.

Another important issue of object detection is how to explore effective training sets as the performance is somehow

largely depends on quantity and quality of both positive and negative samples. Online bootstrapping (or hard negative mining [139]) for CNN training has recently gained interest due to its importance for intelligent cognitive systems interacting with dynamically changing environments [58], [140], [141]. [142] proposes a novel bootstrapping technique called online hard example mining (OHEM) for training detection models based on CNNs. It simplifies the training process by automatically selecting the hard examples. Meanwhile, it only computes the feature maps of an image once, and then forwards all region-of-interests (RoIs) of the image on top of these feature maps. Thus it is able to find the hard examples with a small extra computational cost.

C. Object tracking

The success in object tracking relies heavily on how robust the representation of target appearance is against several challenges such as view point changes, illumination changes, and occlusions. There are several attempts to employ CNNs for visual tracking. Fan *et al.* [143] use CNN as a base learner. It learns a separate class-specific network to track objects. In [143], the authors design a CNN tracker with a shift-variant architecture. Such an architecture plays a key role so that it turns the CNN model from a detector into a tracker. The features are learned during offline training. Different from traditional trackers which only extract local spatial structures, this CNN based tracking method extracts both spatial and temporal structures by considering the images of two consecutive frames. Because the large signals in the temporal information tend to occur near objects that are moving, the temporal structures provide a crude velocity signal to tracking.

Li *et al.* [144] propose a target-specific CNN for object tracking, where the CNN is trained incrementally during tracking with new examples obtained online. They employ a candidate pool of multiple CNNs as a data-driven model of different instances of the target object. Individually, each CNN maintains a specific set of kernels that favourably discriminate object patches from their surrounding background using all available low-level cues. These kernels are updated in an online manner at each frame after being trained with just one instance at the initialization of the corresponding CNN. Instead of learning one complicated and powerful CNN model for all the appearance observations in the past, [144] uses a relatively small number of filters in the CNN within a framework equipped with a temporal adaptation mechanism. Given a frame, the most promising CNNs in the pool are selected to evaluate the hypotheses for the target object. The hypothesis with the highest score is assigned as the current detection window and the selected models are retrained using a warm-start backpropagation which optimizes a structural loss function.

In [145], a CNN object tracking method is proposed to address limitations of handcrafted features and shallow classifier structures in object tracking problem. The discriminative features are first automatically learned via a CNN. To alleviate the tracker drifting problem caused by model update, the

tracker exploits the ground truth appearance information of the object labeled in the initial frames and the image observations obtained online. A heuristic schema is used to judge whether updating the object appearance models or not.

Hong *et al.* [146] propose a visual tracking algorithm based on a pre-trained CNN, where the network is trained originally for large-scale image classification and the learned representation is transferred to describe target. On top of the hidden layers in the CNN, they put an additional layer of an online SVM to learn a target appearance discriminatively against background. The model learned by SVM is used to compute a target-specific saliency map by back-projecting the information relevant to target to input image space. And they exploit the target-specific saliency map to obtain generative target appearance models and perform tracking with understanding of spatial configuration of target.

D. Pose estimation

Since the breakthrough in deep structure learning, many recent works pay more attention to learn multiple levels of representations and abstractions for human-body pose estimation task with CNNs. DeepPose [147] is the first application of CNNs to human pose estimation problem. In this work, pose estimation is formulated as a CNN-based regression problem to body joint coordinates. A cascade of 7-layered CNNs are presented to reason about pose in a holistic manner. Unlike the previous works that usually explicitly design graphical model and part detectors, the DeepPose captures the full context of each body joint by taking the whole image as the input.

Meanwhile, some works exploit CNN to learn representation of local body parts. Ajrun *et al.* [148] present a CNN based end-to-end learning approach for full-body human pose estimation, in which CNN part detectors and an Markov Random Field (MRF)-like spatial model are jointly trained, and pair-wise potentials in the graph are computed using convolutional priors. In a series of papers, Tompson *et al.* [149] use a multi-resolution CNN to compute heat-map for each body part. Different from [148], Tompson *et al.* [149] learn the body part prior model and implicitly the structure of the spatial model. Specifically, they start by connecting every body part to itself and to every other body part in a pair-wise fashion, and use a fully-connected graph to model the spatial prior. As an extension of [149], Tompson *et al.* [51] propose a CNN architecture which includes a position refinement model after a rough pose estimation CNN. This refinement model, which is a Siamese network [150], is jointly trained in cascade with the off-the-shelf model [149]. In a similar work with [149], Chen *et al.* [151], [152] also combine graphical model with CNN. They exploit a CNN to learn conditional probabilities for the presence of parts and their spatial relationships, which are used in unary and pairwise terms of the graphical model. The learned conditional probabilities can be regarded as low-dimensional representations of the body pose. There is also a pose estimation method called dual-source CNN [153] that integrates graphical models and holistic style. It takes the full body image and the holistic view of the local parts as inputs to combine both local and contextual information.

In addition to still image pose estimation with CNN, recently researchers also apply CNN to human pose estimation in videos. Based on the work [149], Jain *et al.* [154] also incorporate RGB features and motion features to a multi-resolution CNN architecture to further improve accuracy. Specifically, The CNN works in a sliding-window manner to perform pose estimation. The input of the CNN is a 3D tensor which consists of an RGB image and its corresponding motion features, and the output is a 3D tensor containing response-maps of the joints. In each response map, the value of each location denote the energy for presence the corresponding joint at that pixel location. The multi-resolution processing is achieved by simply down sampling the inputs and feeding them to the network.

E. Text detection and recognition

The task of recognizing text in image has been widely studied for a long time. Traditionally, optical character recognition (OCR) is the major focus. OCR techniques mainly perform text recognition on images in rather constrained visual environments (*e.g.*, clean background, well-aligned text). Recently, the focus has been shifted to text recognition on scene images due to the growing trend of high-level visual understanding in computer vision research. The scene images are captured in unconstrained environments where there exists a large amount of appearance variations which poses great difficulties to existing OCR techniques. Such a concern can be mitigated by using stronger and richer feature representations such as those learned by CNN models. Along the line of improving the performance of scene text recognition with CNN, a few works have been proposed. The works can be coarsely categorized into three types: (1) text detection and localization without recognition, (2) text recognition on cropped text images, and (3) end-to-end text spotting that integrates both text detection and recognition:

1) *Text detection*: One of the pioneering works to apply CNN for scene text detection is [155]. The CNN model employed by [155] learns on cropped text patches and non-text scene patches to discriminate between the two. The text are then detected on the response maps generated by the CNN filters given the multiscale image pyramid of the input. To reduce the search space for text detection, [156] proposes to obtain a set of character candidates via Maximally Stable Extremal Regions (MSER) and filter the candidates by CNN classification. Another work that combines MSER and CNN for text detection is [157]. In [157], CNN is used to distinguish text-like MSER components from non-text components, and cluttered text components are split by applying CNN in a sliding window manner followed by Non-Maximal Suppression (NMS). Other than localization of text, there is an interesting work [158] that makes use of CNN to determine whether the input image contains text, without telling where the text is exactly located. In [158], text candidates are obtained using MSER which are then passed into a CNN to generate visual features, and lastly the global features of the images are constructed by aggregating the CNN features in a Bag-of-Words (BoW) framework.

2) *Text recognition*: [159] proposes a CNN model with multiple softmax classifiers in its final layer, which is formulated in such a way that each classifier is responsible for character prediction at each sequential location in the multi-digit input image. As an attempt to recognize text without using lexicon and dictionary, [160] introduces a novel Conditional Random Fields (CRF)-like CNN model to jointly learn character sequence prediction and bigram generation for scene text recognition. The more recent text recognition methods supplement conventional CNN models with variants of recurrent neural networks (RNN) to better model the sequence dependencies between characters in text. In [161], CNN extracts rich visual features from character-level image patches obtained via sliding window, and the sequence labelling is carried out by an enhanced RNN variant called Long Short-Term Memory (LSTM) [162]. The method presented in [163] is very similar to [161], except that in [163], lexicon can be taken into consideration to enhance text recognition performance.

3) *End-to-end text spotting*: For end-to-end text spotting, [12] applies a CNN model originally trained for character classification to perform text detection. Going in a similar direction as [12], the CNN model proposed in [164] enables feature sharing across the four different subtasks of an end-to-end text spotting system: text detection, character case-sensitive and insensitive classification, and bigram classification. [165] makes use of CNNs in a very comprehensive way to perform end-to-end text spotting. In [165], the major subtasks of its proposed system, namely text bounding box filtering, text bounding box regression, and text recognition are each tackled by a separate CNN model.

F. Visual saliency detection

The technique to locate important regions in imagery is referred to as visual saliency prediction. It is a challenging research topic, with a vast number of computer vision and image processing applications facilitated by it. Recently, a couple of works have been proposed to harness the strong visual modeling capability of CNNs for visual saliency prediction.

Multi-contextual information is a crucial prior in visual saliency prediction, and it has been used concurrently with CNN in most of the considered works [166–170]. [166] introduces a novel saliency detection algorithm which sequentially exploits local context and global context. The local context is handled by a CNN model which assigns a local saliency value to each pixel given the input of local image patches, while the global context (object-level information) is handled by a deep fully-connected feedforward network. In [167], the CNN parameters are shared between the global-context and local-context models, for predicting the saliency of superpixels found within object proposals. The CNN model adopted in [168] is pre-trained on large-scale image classification dataset and then shared among different contextual levels for feature extraction. The outputs of the CNN at different contextual levels are then concatenated as input to be passed into a trainable fully-connected feedforward network for saliency prediction. Similar to [167], [168], the CNN model used in

[169] for saliency prediction are shared across three CNN streams, with each stream taking input of a different contextual scale. [170] derives a spatial kernel and a range kernel to produce two meaningful sequences as 1-D CNN inputs, to describe color uniqueness and color distribution respectively. The proposed sequences are advantageous over inputs of raw image pixels because they can reduce the training complexity of CNN, while being able to encode the contextual information among superpixels.

There are also CNN-based saliency prediction approaches [171–173] that do not consider multi-contextual information. Instead, they rely very much on the powerful representation capability of CNN. In [171], an ensemble of CNNs is derived from a large number of randomly instantiated CNN models, to generate good features for saliency detection. The CNN models instantiated in [171] are however not deep enough because the maximum number of layers is capped at three. By using a pre-trained and deeper CNN model with 5 convolutional layers, [172] (Deep Gaze) learns a separate saliency model to jointly combine the responses from every CNN layer and predict saliency values. [173] is the only work making use of CNN to perform visual saliency prediction in an end-to-end manner, which means the CNN model accepts raw pixels as input and produces saliency map as output. [173] argues that the success of the proposed end-to-end method is attributed to its not-so-deep CNN architecture which attempts to prevent overfitting.

G. Action recognition

Action recognition, the behaviour analysis of human subjects and classifying their activities based on their visual appearance and motion dynamics, is one of the challenging problems in computer vision. Generally, this problem can be divided to two major groups: action analysis in still images and in videos. For both of these two groups, effective CNN based methods are proposed. In this subsection we briefly introduce the latest advances on these two groups.

1) *Action Recognition in Still Images*: The work of [174] has shown the output of last few layers of a trained CNN can be used as a general visual feature descriptor for a variety of tasks. The same intuition is utilized for action recognition by [8], [175], in which they use the outputs of the penultimate layer of a pre-trained CNN to represent full images of actions as well as the human bounding boxes inside them, and achieve a high level of performance in action classification. Gkioxari *et al.* [176] add a part detection to this framework. Their part detector is a CNN based extension to the original Poselet [177] method.

CNN based representation of contextual information is utilized for action recognition in [178]. They search for the most representative secondary region within a large number of object proposal regions in the image and add contextual features to the description of the primary region (ground truth bounding box of human subject) in a bottom-up manner. They utilize a CNN to represent and fine-tune the representations of the primary and the contextual regions.

2) *Action Recognition in Video Sequences*: Applying CNNs on videos is challenging because traditional CNNs are designed to represent two dimensional pure spatial signals but in videos a new temporal axis is added which is essentially different from the spatial variations in images. The sizes of the video signals are also in higher orders in comparison to those of images which makes it more difficult to apply convolutional networks on. Ji *et al.* [179] propose to consider the temporal axis in a similar manner as other spatial axes and introduce a network of 3D convolutional layers to be applied on video inputs. Recently Tran *et al.* [180] study the performance, efficiency, and effectiveness of this approach and show its strengths compared to other approaches.

Another approach to apply CNNs on videos is to keep the convolutions in 2D and fuse the feature maps of consecutive frames, as proposed by [181]. They evaluate three different fusion policies: late fusion, early fusion, and slow fusion, and compare them with applying the CNN on individual single frames. One more step forward for better action recognition via CNNs is to separate the representation to spatial and temporal variations and train individual CNNs for each of them, as proposed by Simonyan and Zisserman [182]. First stream of this framework is a traditional CNN applied on all the frames and the second receives the dense optical flow of the input videos and trains another CNN which is identical to the spatial stream in size and structure. The output of the two streams are combined in a class score fusion step. Chéron *et al.* [183] utilize the two stream CNN on the localized parts of the human body and show the aggregation of part-based local CNN descriptors can effectively improve the performance of action recognition. Another approach to model the dynamics of videos differently from spatial variations, is to feed the CNN based features of individual frames to a sequence learning module *e.g.*, a recurrent neural network. [184] studies different configurations of applying LSTM units as the sequence learner in this framework.

H. Scene Labeling

Scene labeling aims to relate one semantic class (road, water, sea etc.) to each pixel of the input image. CNNs are used to model the class likelihood of pixels directly from local image patches. They are able to learn strong features and classifiers to discriminate the local visual subtleties. Farabet *et al.* [185] have pioneered to apply CNNs to scene labeling tasks. They feed their Multi-scale ConvNet with different scale image patches, and they show that the learned network is able to perform much better than systems with hand-crafted features. Besides, this network is also successfully applied to RGB-D scene labeling [186]. To enable the CNNs to have a large field of view over pixels, Pinheiro *et al.* [187] develop the recurrent CNNs. More specifically, the identical CNNs are applied recurrently to the output maps of CNNs in the previous iterations. By doing this, they can achieve slightly better labeling results while significantly reduces the inference times. Shuai *et al.* [188–190] train the parametric CNNs by sampling image patches, which speeds up the training time dramatically. They find that patch-based CNNs suffer from local ambiguity

problems, and [188] solve it by integrating global beliefs. [189] and [190] use the recurrent neural networks to model the contextual dependencies among image features from CNNs, and dramatically boost the labeling performance.

Meanwhile, researchers are exploiting to use the pre-trained deep CNNs for object semantic segmentation. Mostajabi *et al.* [191] apply the local and proximal features from a ConvNet and apply the Alex-net [6] to obtain the distant and global features, and their concatenation gives rise to the zoom-out features. They achieve very competitive results on the semantic segmentation tasks. Long *et al.* [134] train a fully convolutional Network (FCN) to directly predict the input images to dense label maps. The convolution layers of the FCNs are initialized from the model pre-trained on ImageNet classification dataset, and the deconvolution layers are learned to upsample the resolution of label maps. Chen *et al.* [192] also apply the pre-trained deep CNNs to emit the labels of pixels. Considering that the imperfectness of boundary alignment, they further use fully connected Conditional Random Fields (CRF) to boost the labeling performance.

VI. CONCLUSIONS AND OUTLOOK

Deep CNNs have made breakthroughs in processing image, video, speech and text. In this paper, we give an extensive survey on recent advances of CNNs mainly from the computer vision perspective. We have discussed the improvements of CNN on different aspects, namely, layer design, activation function, loss function, regularization, optimization and fast computation. Beyond surveying the advances of each aspect of CNN, we also introduce the application of CNN on computer vision tasks, including image classification, object detection, object tracking, pose estimation, text detection, visual saliency detection, action recognition and scene labeling.

Although CNNs have achieved great success in experimental evaluations, there are still lots of works deserve further investigation. Firstly, since the recent CNNs are becoming deeper and deeper, they require large scale dataset and massive computing power for training. Manually collecting labelled dataset requires huge amounts of human efforts. Thus, it is desired to explore unsupervised learning of CNNs. Meanwhile, to speedup training procedure, although there are already some asynchronous SGD algorithms have shown promising result by using CPU and GPU clusters, it is still worth to develop effective and scalable parallel training algorithms. At testing time, these deep models are high memory demanding and time consuming, which makes them not suitable to be deployed on mobile platforms that have limited resources. It is important to investigate how to reduce the complexity and obtain fast-to-execute models without loss of accuracy.

Secondly, we note that one major barrier of applying CNN on new task is that it requires considerable skill and experience to select suitable hyper-parameters such as the learning rate, kernel sizes of convolutional filters, the number of layers etc. These hyper-parameters have internal dependencies which makes them especially expensive for tuning. Recent works have shown that there is vast room to improve current optimization techniques for learning deep CNN architectures [10], [23], [193].

Finally, the solid theory of CNNs is still lacking. Current CNN model works very well on various applications. However, we do not even know why and how it works essentially. It is desirable to make more efforts on investigating the fundamental principles of CNNs. Meanwhile, it is also worth exploring how to leverage natural visual perception mechanism to further improve the design of CNN. We hope that this paper does not only provide a better understanding of CNNs but also facilitates future research activities and application developments in the field of CNNs.

ACKNOWLEDGMENT

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Prime Ministers Office, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office.

REFERENCES

- [1] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, 1968.
- [2] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, 1980.
- [3] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *NIPS*. Citeseer, 1990.
- [4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.
- [5] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *IJCNN*, 1989.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
- [7] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.
- [8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
- [9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, 2014.
- [10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
- [11] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.
- [12] T. Wang, D. Wu, A. Coates, and A. Ng, "End-to-end text recognition with convolutional neural networks," in *ICPR*, 2012.
- [13] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009.
- [14] Y. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *ICML*, 2010.
- [15] M. Ranzato, F. J. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *CVPR*, 2007.
- [16] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.
- [17] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.
- [18] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT*, 2010.
- [19] R. G. J. Wijnhoven and P. H. N. de With, "Fast training of object detection using stochastic gradient descent," in *ICPR*, 2010.
- [20] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, "Parallelized stochastic gradient descent," in *NIPS*, 2010.
- [21] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

- [22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Re-thinking the Inception Architecture for Computer Vision," 2015, arXiv:1512.00567v1.
- [23] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.
- [24] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area mt," *Vision research*, 1998.
- [25] A. Hyvärinen and U. Köster, "Complex cell pooling and the statistics of natural images," *Network: Computation in Neural Systems*, 2007.
- [26] J. B. Estrach, A. Szlam, and Y. Lecun, "Signal recovery from pooling representations," in *ICML*, 2014.
- [27] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," in *Machine Learning and Knowledge Discovery in Databases*, 2014.
- [28] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropout," in *ICML*, 2013.
- [29] D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *Rough Sets and Knowledge Technology*, 2014.
- [30] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," *CoRR*, 2013.
- [31] O. Rippel, J. Snoek, and R. P. Adams, "Spectral representations for convolutional neural networks," *arXiv preprint arXiv:1506.03767*, 2015.
- [32] M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through ffts," *arXiv preprint arXiv:1312.5851*, 2013.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *ECCV*, 2014.
- [34] S. Singh, A. Gupta, and A. Efros, "Unsupervised discovery of mid-level discriminative patches," *ECCV*, 2012.
- [35] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, 2014.
- [36] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *PAMI*, 2012.
- [37] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML*, 2013.
- [38] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011.
- [39] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean *et al.*, "On rectified linear units for speech processing," in *ICASSP*. IEEE, 2013.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *arXiv preprint arXiv:1502.01852*, 2015.
- [41] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.
- [42] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.
- [43] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *arXiv preprint arXiv:1302.4389*, 2013.
- [44] J. T. Springenberg and M. Riedmiller, "Improving deep neural networks with probabilistic maxout units," *arXiv preprint arXiv:1312.6116*, 2013.
- [45] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *ICML*, 2004.
- [46] Y. LeCun, C. Cortes, and C. J. Burges, "The mnist database of handwritten digits," 1998.
- [47] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, 2005.
- [48] J. Lin, O. Morere, V. Chandrasekhar, A. Veillard, and H. Goh, "Deephash: Getting regularization, depth and fine-tuning right," *arXiv preprint arXiv:1501.04711*, 2015.
- [49] S. Wang and C. Manning, "Fast dropout training," in *ICML*, 2013.
- [50] J. Ba and B. Frey, "Adaptive dropout for training deep neural networks," in *NIPS*, 2013.
- [51] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," *arXiv preprint arXiv:1411.4280*, 2014.
- [52] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," *arXiv preprint arXiv:1412.0233*, 2014.
- [53] D. Mishkin and J. Matas, "All you need is a good init," *arXiv preprint arXiv:1511.06422*, 2015.
- [54] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *ICML*, 2013.
- [55] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010.
- [56] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.
- [57] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *ICCV*, 2015.
- [58] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *ICCV*, 2015.
- [59] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in *ICCV*, 2015.
- [60] G. B. Orr and K.-R. Müller, *Neural networks: tricks of the trade*, 2003.
- [61] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, 1999.
- [62] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.
- [63] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *JMLR*, 2011.
- [64] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [65] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in *NIPS*, 2011.
- [66] Y. Bengio, "Deep learning of representations: Looking forward," in *Statistical Language and Speech Processing*, 2013.
- [67] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le *et al.*, "Large scale distributed deep networks," in *NIPS*, 2012.
- [68] T. Paine, H. Jin, J. Yang, Z. Lin, and T. Huang, "Gpu asynchronous stochastic gradient descent to speed up neural network training," *arXiv preprint arXiv:1312.6186*, 2013.
- [69] Y. Zhuang, W.-S. Chin, Y.-C. Juan, and C.-J. Lin, "A fast parallel sgd for matrix factorization in shared memory systems," in *RecSys*. ACM, 2013.
- [70] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [71] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.
- [72] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *NIPS*, 2015.
- [73] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *arXiv preprint arXiv:1603.05027*, 2016.
- [74] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.
- [75] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cudnn: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, 2014.
- [76] N. Vasilache, J. Johnson, M. Mathieu, S. Chintala, S. Piantino, and Y. LeCun, "Fast convolutional nets with fbfft: A gpu performance evaluation," *arXiv preprint arXiv:1412.7580*, 2014.
- [77] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, 2013.
- [78] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *ICASSP*, 2013.
- [79] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition," in *INTER-SPEECH*, 2013, pp. 2365–2369.
- [80] M. Denil, B. Shakibi, L. Dinh, N. de Freitas *et al.*, "Predicting parameters in deep learning," in *NIPS*, 2013.
- [81] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in *NIPS*, 2014.
- [82] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding up convolutional neural networks with low rank expansions," *arXiv preprint arXiv:1405.3866*, 2014.
- [83] G. J. Sullivan, "Efficient scalar quantization of exponential and laplacian random variables," *Information Theory, IEEE Transactions on*, 1996.
- [84] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," *arXiv preprint arXiv:1412.6115*, 2014.

- [85] Y. Chen, T. Guan, and C. Wang, "Approximate nearest neighbor search by residual vector quantization," *Sensors*, 2010.
- [86] R. Balasubramanian, C. A. Bouman, and J. P. Allebach, "Sequential scalar quantization of color images," *Journal of Electronic Imaging*, 1994.
- [87] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *Neural Networks, IEEE Transactions on*, 1997.
- [88] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *null*. IEEE, 2003.
- [89] F. J. Huang and Y. LeCun, "Large-scale learning with svm and convolutional for generic object categorization," in *CVPR*, 2006.
- [90] D. C. Cireşan, U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *IJCAI*, 2011.
- [91] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.
- [92] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *IJCV*, 2014.
- [93] A.-M. Tousch, S. Herbin, and J.-Y. Audibert, "Semantic hierarchies for image annotation: A survey," *Pattern Recognition*, 2012.
- [94] N. Srivastava and R. R. Salakhutdinov, "Discriminative transfer learning with tree-based priors," in *NIPS*, 2013.
- [95] Z. Wang, X. Wang, and G. Wang, "Learning fine-grained features via a cnn tree for large-scale classification," *arXiv preprint arXiv:1511.04534*, 2015.
- [96] T. Xiao, J. Zhang, K. Yang, Y. Peng, and Z. Zhang, "Error-driven incremental learning in deep convolutional neural network for large-scale image classification," in *ACMMM*, 2014.
- [97] Z. Yan, V. Jagadeesh, D. DeCoste, W. Di, and R. Piramuthu, "Hd-cnn: Hierarchical deep convolutional neural network for image classification," *arXiv preprint arXiv:1410.0736*, 2014.
- [98] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *ICVGIP*, 2008.
- [99] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.
- [100] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *CVPR*, 2014.
- [101] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *CVPR*, 2011.
- [102] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *CVPR*, 2015.
- [103] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *CVPR*, 2014.
- [104] S. Branson, G. Van Horn, P. Perona, and S. Belongie, "Improved bird species recognition using pose normalized deep convolutional nets," in *BMVC*, 2014.
- [105] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *ECCV*, 2014.
- [106] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, 2013.
- [107] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep lac: Deep localization, alignment and classification for fine-grained recognition," in *CVPR*, 2015.
- [108] J. P. Pluim, J. A. Maintz, M. Viergever *et al.*, "Mutual-information-based registration of medical images: a survey," *T-MI*, 2003.
- [109] J. Krause, T. Gebru, J. Deng, L.-J. Li, and L. Fei-Fei, "Learning features and parts for fine-grained recognition," in *ICPR*, 2014.
- [110] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *CVPR*, 2015.
- [111] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *CVPR*, 2015.
- [112] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," *arXiv preprint arXiv:1504.07889*, 2015.
- [113] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *JSAI*, 1999.
- [114] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.
- [115] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.
- [116] O. Matan, H. S. Baird, J. Bromley, C. J. Burges, J. S. Denker, L. D. Jackel, Y. Le Cun, E. P. Pednault, W. D. Satterfield, C. E. Stenard *et al.*, "Reading handwritten digits: A zip code recognition system," *Computer*, 1992.
- [117] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *PAMI*, 1998.
- [118] S. J. Nowlan and J. C. Platt, "A convolutional neural network hand tracker," *NIPS*, 1995.
- [119] C. Garcia and M. Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," *PAMI*, 2004.
- [120] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *NIPS*, 2013.
- [121] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *CVPR*, 2014.
- [122] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov, "Scalable, high-quality object detection," *arXiv preprint arXiv:1412.1441*, 2014.
- [123] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," *arXiv preprint arXiv:1409.5403*, 2014.
- [124] P.-A. Savalle, S. Tsogkas, G. Papandreou, and I. Kokkinos, "Deformable part models with cnn features," in *ECCV*, 2014.
- [125] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *CVPR*, 2013.
- [126] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian *et al.*, "Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection," *arXiv preprint arXiv:1409.3505*, 2014.
- [127] R. Vaillant, C. Monrocq, and Y. Le Cun, "Original approach for the localisation of objects in images," in *Vision, Image and Signal Processing, IEE Proceedings-*, 1994.
- [128] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
- [129] I. Endres and D. Hoiem, "Category independent object proposals," in *ECCV*, 2010.
- [130] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *PAMI*, 2012.
- [131] J. Carreira and C. Sminchisescu, "Cpmc: Automatic object segmentation using constrained parametric min-cuts," *PAMI*, 2012.
- [132] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.
- [133] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
- [134] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.
- [135] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *PAMI*, 2015.
- [136] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
- [137] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *ICCV*, 2015.
- [138] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. So Kweon, "Attentionnet: Aggregating weak directions for accurate object detection," in *CVPR*, 2015.
- [139] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, 2010.
- [140] I. Loshchilov and F. Hutter, "Online batch selection for faster training of neural networks," *arXiv preprint arXiv:1511.06343*, 2015.
- [141] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, and F. Moreno-Noguer, "Fracking deep convolutional image descriptors," *arXiv preprint arXiv:1412.6537*, 2014.
- [142] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," *arXiv preprint arXiv:1604.03540*, 2016.
- [143] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *Neural Networks, IEEE Transactions on*, 2010.
- [144] H. Li, Y. Li, and F. Porikli, "Deeptrack: Learning discriminative feature representations by convolutional neural networks for visual tracking," in *BMVC*, 2014.
- [145] Y. Chen, X. Yang, B. Zhong, S. Pan, D. Chen, and H. Zhang, "Cn-tracker: Online discriminative object tracking via deep convolutional neural network," *Applied Soft Computing*, 2015.
- [146] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," *arXiv preprint arXiv:1502.06796*, 2015.

- [147] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *CVPR*, 2014.
- [148] "Learning human pose estimation features with convolutional networks," 2014.
- [149] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *NIPS*, 2014.
- [150] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," *IJPRAI*, 1993.
- [151] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *NIPS*, 2014.
- [152] X. Chen and A. Yuille, "Parsing occluded people by flexible compositions," in *CVPR*, 2015.
- [153] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *CVPR*, 2015.
- [154] A. Jain, J. Tompson, Y. LeCun, and C. Bregler, "Modeep: A deep learning framework using motion features for human pose estimation," in *ACCV*, 2014.
- [155] M. Delakis and C. Garcia, "Text detection with convolutional neural networks," in *VISAPP*, 2008.
- [156] H. Xu and F. Su, "Robust seed localization and growing with deep convolutional features for scene text detection," in *ICMR*, 2015.
- [157] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced msr trees," in *ECCV*, 2014.
- [158] C. Zhang, C. Yao, B. Shi, and X. Bai, "Automatic discrimination of text and non-text natural images," in *ICDAR*, 2015.
- [159] I. J. Goodfellow, J. Ibarz, S. Arnaud, and V. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks," in *ICLR*, 2014.
- [160] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep structured output learning for unconstrained text recognition," in *ICLR*, 2015.
- [161] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, "Reading scene text in deep convolutional sequences," *CoRR*, 2015.
- [162] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, 2000.
- [163] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *CoRR*, 2015.
- [164] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *ECCV*, 2014.
- [165] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *IJCV*, 2015.
- [166] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *CVPR*, 2015.
- [167] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015.
- [168] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015.
- [169] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *CVPR*, 2015.
- [170] S. He, R. W. Lau, W. Liu, Z. Huang, and Q. Yang, "Supercnn: A superpixelwise convolutional neural network for salient object detection," *IJCV*, 2015.
- [171] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *CVPR*, 2014.
- [172] M. Krmmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," in *ICLR*, 2015.
- [173] J. Pan and X. Gir-i Nieto, "End-to-end convolutional network for saliency prediction," in *CVPR*, 2015.
- [174] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," 2014.
- [175] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *CVPR*, 2014.
- [176] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," 2014.
- [177] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *CVPR*, 2013.
- [178] G. Gkioxari, R. B. Girshick, and J. Malik, "Contextual action recognition with r*cnn," *CoRR*, 2015.
- [179] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *PAMI*, 2013.
- [180] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: generic features for video analysis," *CoRR*, 2014.
- [181] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.
- [182] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., 2014.
- [183] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: pose-based CNN features for action recognition," *CoRR*, 2015.
- [184] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," June 2015.
- [185] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *PAMI*, 2013.
- [186] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.
- [187] P. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *ICML*, 2014.
- [188] B. Shuai, G. Wang, Z. Zuo, B. Wang, and L. Zhao, "Integrating parametric and non-parametric models for scene labeling," in *CVPR*, 2015.
- [189] B. Shuai, Z. Zuo, and W. Gang, "Quaddirectional 2d-recurrent neural networks for image labeling," *Signal Processing Letters*, 2015.
- [190] B. Shuai, Z. Zuo, G. Wang, and B. Wang, "Dag-recurrent neural networks for scene labeling," *arXiv preprint arXiv:1509.00552*, 2015.
- [191] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *CVPR*, 2015.
- [192] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.
- [193] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," *arXiv preprint arXiv:1603.09382*, 2016.